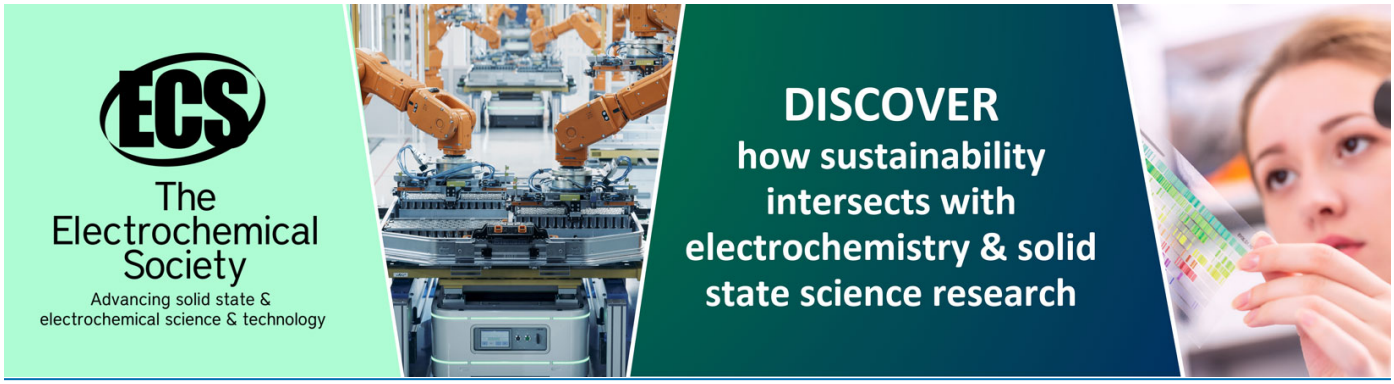
Inverse problem of HIV cell dynamics using Genetic Algorithms



Inverse problem of HIV cell dynamics using Genetic Algorithms

J A González, F S Guzmán
Laboratorio de Inteligencia Artificial y Supercómputo. Instituto de Física y Matemáticas, Universidad Michoacana de San Nicolás de Hidalgo, Morelia México

E-mail: gonzalez@ifm.umich.mx, guzman@ifm.umich.mx

Abstract. In order to describe the cell dynamics of T-cells in a patient infected with HIV, we use a flavour of Perelson's model. This is a non-linear system of Ordinary Differential Equations that describes the evolution of healthy, latently infected, infected T-cell concentrations and the free viral cells. Different parameters in the equations give different dynamics. Considering the concentration of these types of cells is known for a particular patient, the inverse problem consists in estimating the parameters in the model. We solve this inverse problem using a Genetic Algorithm (GA) that minimizes the error between the solutions of the model and the data from the patient. These errors depend on the parameters of the GA, like mutation rate and population, although a detailed analysis of this dependence will be described elsewhere.

1. Introduction

Modelling the dynamics of HIV and AIDS is important because this disease represents a main concern in public health policies, as infected people are counted in dozens of millions around the world [1]. There are two important concerns: prevention and treatment. The treatment involves a constant monitoring of each patient, specifically the dynamics of infected cell concentrations in the blood and is the subject we focus on in this manuscript.

From the clinical point of view, what is required for the present model is that an essential component of the immune system are the Lymphocytes which destroy invaders. Lymphocytes are of two types B-cells and T-cells, being B-cells antibody factories producing antibodies as fast as they can and also clone themselves, whereas T-cells either direct the activity of B-cells (called CD4+T-cells) or act as suppressors (called CD8+T-cells) destroying infected cells and thus damp out the activity of the immune system. The AIDS has three stages, one including the initial infection, a second one of latency and the third one corresponding to a runaway destruction of the immune system. During the latency lapse healthy T-cells are infected although its number remains high. When the concentration of T-cells decreases and that of HIV virus cells increases, it is the AIDS stage.

In practice, as an example, cytometry is a procedure useful to count the concentration of healthy, latently infected and infected T-cells in the blood stream, for instance using the dispersion of laser light, that depends on the enzymes covering the cells [2]. Knowing the concentration of CD4+T-cells it is possible to wonder about modelling the evolution of the different cell populations. Among the various models trying to describe the dynamics of CD4+T-cells (see for instance [3,4]), a simple one describing the phases of latency and the destruction of the immune system is Perelson's model [5,6,7].

In general, a model is expected to describe clinical data and make predictions. Starting from patients data we define the inverse problem of determining the parameters of the model. Knowing the parameters allows one to determine the evolution of the concentration of healthy and infected CD4+T-cells, so as possibly infer the conditions during an early stage of the disease.

Assuming Perelson's model, we intend to solve the inverse problem and reconstruct the coefficients of the evolution model for the cell populations, out of a time series that in a realistic case might correspond to the data taken from a particular patient. This means, we assume that there is a Perelson system of equations for each patient with different parameters. In this sense, what we propose is a personalized set of equations that can be used to diagnose and decide on the treatment of each particular patient. This paper is about testing a method to estimate the parameters characterizing a

given patient. Our method to solve this inverse problem uses Genetic Algorithms (GAs) to find the best combination of parameters that minimizes the difference between the experimental data (in our case generated numerically) and the numerical solution to the Perelson's model.

The paper is organized as follows. In Section 2 we describe the particular model of T-cell populations, whereas in Section 3 we present the GAs applied to estimate the parameters of the ODE system associated to the model. In Section 4 we present some results and in 5 draw some final comments.

2. The T-cell dynamics model

Among the various models describing the evolution of T-cell in the blood stream in patients, in order to show how our method works, we use the basic Perelson model [5]. Following [8], the variables that evolve under this approach are the healthy T-cells (X), latently infected T-cells (Y), actively infected T-cells (Z) and free viral cells (W), which obey the following set of non-linear coupled ODEs

$$\begin{aligned}
\frac{dX}{dt} &= s + rX\left(1 - \frac{X+Y+Z}{x_{\max}}\right) - \mu_X X - k_1 XW, \\
\frac{dY}{dt} &= k_1 XW - \mu_Y Y - k_2 Y, \\
\frac{dZ}{dt} &= k_2 Y - \mu_Z Z, \\
\frac{dW}{dt} &= N\mu_Z Z - k_1 XW - \mu_W W
\end{aligned} \tag{1}$$

Here s is the rate at which T-cells enter the blood stream; r grades the logistic growth rate in the number of CD4+T cells, and their evolution as a function of actively infected and latently infected cells; the concentration of destroyed T-cells is proportional to the population of infected cells with clearance k_1; the coefficient $-\mu_X$ is based on how patients recover after therapy; $x_{\max}$ is the maximum possible CD4+T concentration; μ_Y is the death rate of latently infected CD4+T cells; μ_Z is the death rate of actively infected CD4+T population, μ_W is the death rate of free virus cells, k_1 is the rate of infection of CD4+T by the free virus, k_2 is the rate at which latently infected CD4+T cells convert to actively infected cells and N is the number of free virus produced by lysing a CD4+T cell.

Solving system (1) is straightforward provided initial conditions and specific values of the coefficients/parameters. We solve these equations using a 4th order accurate Runge-Kutta integrator and graphically the solution is shown in Figure 1 for the parameters in Table 1, that we set following [8]. In this case the solution approaches asymptotic values in all variables.

In order to test our approach, this solution will work as our experimental data, consisting of the four time series of $(\tilde{X},\tilde{Y},\tilde{Z},\tilde{W})$ which in a realistic case would be the actual measurements from a patient's blood stream.

variable	value	units	variable	value	units
s	10	$mm^{-3}day^{-1}$	r	0.03	day^{-1}
$x_{\max}$	1500	mm^{-3}	μ_X	0.02	day^{-1}
μ_Y	0.02	day^{-1}	μ_Z	0.24	day^{-1}
μ_W	2.4	day^{-1}	k_1	.000024	$mm^{-3}day^{-1}$
k_2	0.003	day^{-1}	N	500	
X_0	500	mm^{-3}	Y_0	0	mm^{-3}
Z_0	0	mm^{-3}	W_0	0.001	mm^{-3}

Tabla 1. Parameters used to construct an illustrative solution. Such solution will also be considered as data from a patient, being these parameters the target to be found by the GA. X_0, Y_0, Z_0, W_0 are the initial conditions of the four variables.

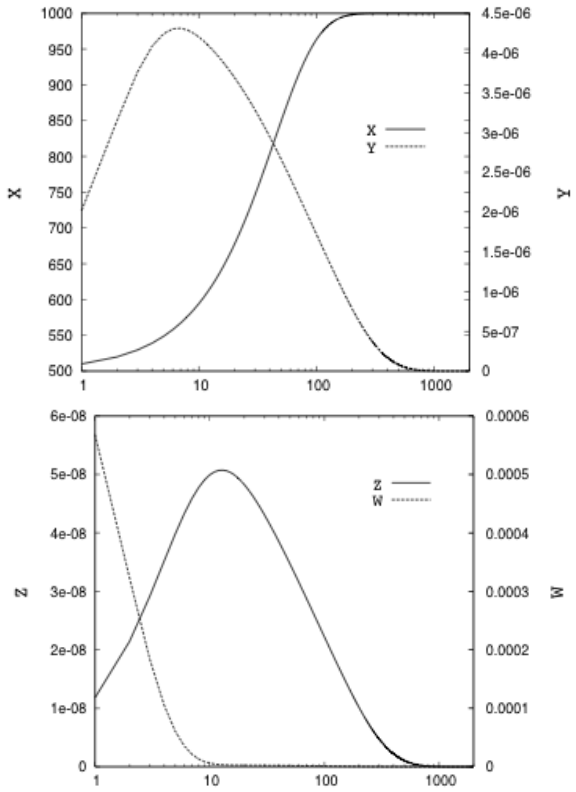


Figure 1. Integration of the system during 2000 days for the parameters in Table 1. All the variables reach a stationary state within this time domain. The horizontal axes are in days.

3. Parameter estimates using Genetic Algorithms

The initial conditions and the various parameters of system (1) are patient dependent. Counts of the four variables before or during treatment work as experimental data that should be fit by the solutions of the system. Thus, the parameters in these equations are particular of each patient submitted to his own treatment. In order to foresee the evolution of the patient as a function of experimental data it is necessary to estimate the whole set of coefficients in system (1). In this manuscript we present an efficient method that estimates these parameters using GAs.

As mentioned before, instead of using realistic data from patients to test our method, we consider the solution of the system for the parameters in Table 1 as our time series data. The advantage of this approach is that we know the values of the parameters and we can measure the accuracy in the parameter estimate of our GAs.

Inverse problem strategy. We assume each patient is described by a Perelson system of equations with unknown coefficients. Therefore we set the inverse problem as an optimization problem where the variables are the coefficients of system (1). In order to solve this problem we use a Genetic Algorithm described as follows. We assume that the array of numbers $\{s, r, x_{\max}, \mu_X, \mu_Y, \mu_Z, \mu_W, k_1, k_2, N\}$ is the DNA of an organism, and each entrance is a gene. With this assumption the GA contains the following elements

- Initial Population. We define an initial population of $j = 1, ..., N_{org}$ organisms with random entrances with the values of the parameters of the data, but randomly shifted within a given range $0 < \beta < 1$. For

example, we chose the parameter s to be $s=10(1\pm\beta\varepsilon)$, with ε a random number between 0 and 1. N_{org} is the size of the population of organisms we start the GA with.

- For each organism j, a particular system of equations of type (1) is defined and is solved numerically.
- The resulting solution for each organism is compared with the data $(\tilde{X},\tilde{Y},\tilde{Z},\tilde{W})$.
- We calculate a weighted $L_1(E_{t,j})$ norm of the error between the solution and the time-series for organism j. Considering each of the variables X,Y,Z,W has values within different scales, the error is measured as follows:

$$L_1(E_{t,j})=\sum_i\left(\frac{|X_j-\tilde{X}_j|}{\tilde{X}_j}+\frac{|Y_j-\tilde{Y}_j|}{\tilde{Y}_j}+\frac{|Z_j-\tilde{Z}_j|}{\tilde{Z}_j}+\frac{|W_j-\tilde{W}_j|}{\tilde{W}_j}\right)$$

where i labels a summation over the total number of days in the time series. We define the fitness of the organism j as $F_j=1/L_1(E_{t,j})$ and classify the organisms from the best to the worst fitted.

- The selection: Using a roulette selection algorithm [9,10], we choose a couple of organisms, one called the mother and the other the father, with the only restriction that they cannot be the same. Each couple will generate a new organism using the crossover mechanism described below. We select $N_{org}-2$ couples to create the same number of children that will populate the next generation. The remaining 2 organisms missing to complete the next generation will be the copies of the 2 best fitted DNAs of the current generation.
- The crossover: We choose randomly 5 genes out of the 10 of the mother. Then complete the DNA of a child with the 5 genes from the father not chosen from the mother.
- Mutation: Every individual has 40% probability to mutate. Every gene of a selected individual is forced to mutate within $\alpha\%$ of its current value.

This process is repeated in a non-standard way. Consider that t^k labels the time in the time series (in days) and that $\tilde{X}^k,\tilde{Y}^k,\tilde{Z}^k,\tilde{W}^k$ are the data at t^k. The GA starts at t^0 and evolves the necessary number of generations to find the DNA that produces a numerical solution X^1,Y^1,Z^1,W^1 with coefficients $\{s_1,r_1,x_{\max,1},\mu_{X,1},\mu_{Y,1},\mu_{Z,1},\mu_{W,1},k_{1,1},k_{2,1},N_1\}$, with an error within a given tolerance θ from the data. These coefficients are then used as the initial guess to evolve another number of generations to fit the variables at t^2 and so on and so forth.

4. Results

We thus use the GA to evolve the initially randomly generated organisms within an initial departure β. In Fig. 2 we show the solution using the DNA of the fittest organism found after a number of generation using $\theta=0.1$ and two different values of $\beta=0.1,1$. Notice that the values of β correspond to parameters possibly at 10 and 100% distant from the parameters used to generate the data.

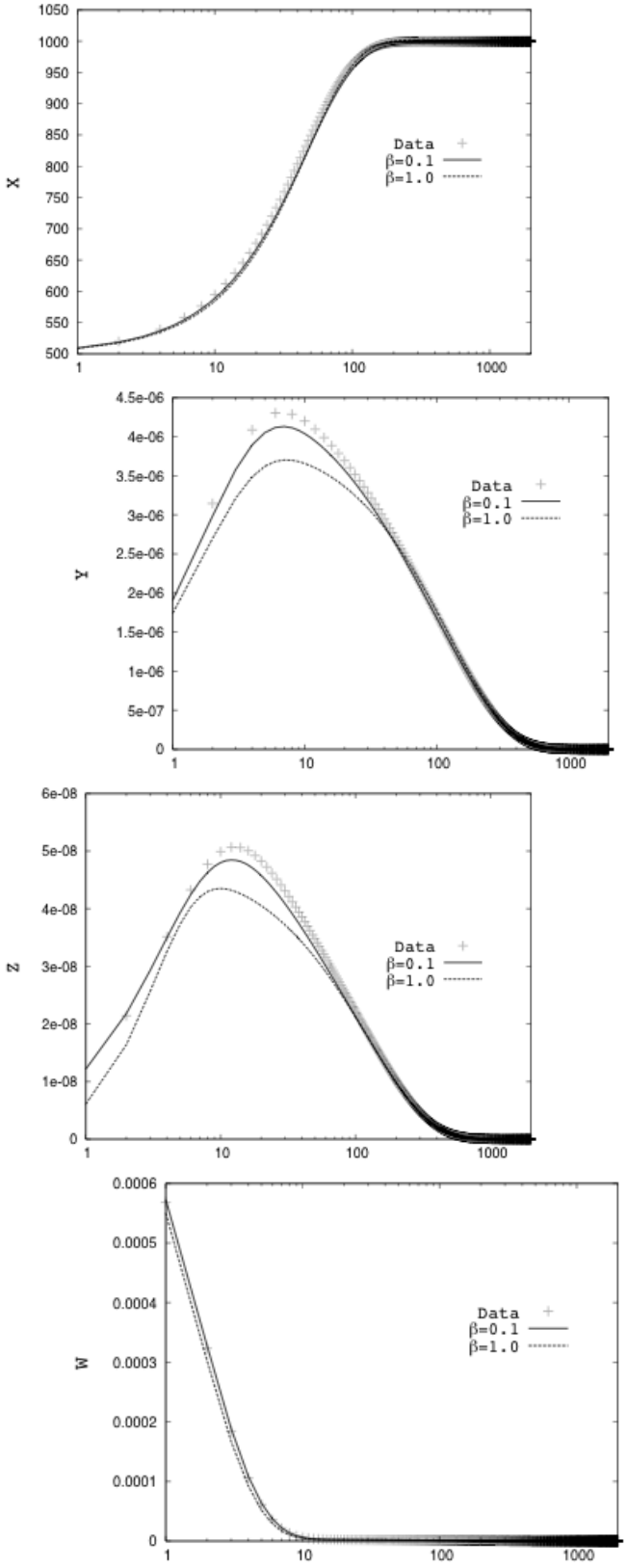

X
Y
Z
W
Data
β=0.1
β=1.0

Figure 2. Solution calculated using the coefficients corresponding to the best fitted organism of a given generation with $\beta = 0.1,1$ and tolerance $\theta = 1$. These are compared with the data from Figure 1. The horizontal axes are in days.

5. Final comments and conclusions

We implemented an application that uses GAs to estimate the parameters of a Perelson's model. This means we solve the inverse problem which in practice implies that, within a given tolerance, we have a method to assign a particular set of equations to a particular patient, based on the T-cell concentration measurements. For this we do not use patient's real data, but a numerical solution, and the GA tracks down the parameters these data were constructed with.

A simple improvement of our method can include the initial conditions of the four variables as genes of the DNA of our organisms. This would allow to estimate X, Y, Z, W during the latency period and possibly bound (going backward until the initial conditions) a lapse of infection. It is believed that early phases of HIV contain crucial information about the further immune response and viral dynamics and is expected to influence the progression of AIDS [11].

Our method works for this non-linear system of equations and is expected to work for straightforward generalization of Perelson's model as well. The complexity of this simple model will allow us to engage the parameter estimation of more elaborate problems, for instance the patient's parameter reconstruction based on the response to treatment considering interrupted therapy [11,12,13,14].

6. Acknowledgments

This research is partly supported by grants CIC-UMSNH-4.9, CIC-UMSNH-4.23 and CONACyT 258726 (Fondo Sectorial de Investigación para la Educación).

7. References

[1] S Armstrong, C Fontaine and A Wilson, Report on the global AIDS epidemic UNAIDS/Joint *United Nations Programme on HIV/AIDS*, (Geneva, Switzerland. Online: http://www.unaids.org)
[2] M G Ormerod 2008 *Flow Cytometry* (Garland Pub.)
[3] J D Murray 2002 *Mathematical Biology I. An Introduction.* (Ed. Springer)
[4] A S Perelson and G Weisbuch 1997 Immunology for physicists. *Rev. Mod. Phys.* **69** 1219-1267
[5] A S Perelson, D E Kirschner, R De Boer 1993 Dynamics of HIV infection of CD4+ T cells. *Mathematical Biosciences* **114** 81-125
[6] A S Perelson and P W Nelson 1999 Mathematical analysis of HIV-1 dynamics in vivo. *SIAM review* **41** 3-44
[7] A S Perelson 2002 Modelling viral and immune system dynamics *Nature Reviews Immunology* **2** 28-36
[8] P A Isihara 2005 Immunological and Epidemiological HIV/AIDS modelling. *THE UMAP JOURNAL **26** 49-90*
[9] J H Holland 1975 *Adaptation in Natural and Artificial Systems* (University of Michigan Press. 2nd Ed. MIT press)
[10] M Mitchell 1998 *An introduction to Genetic Algorithms*. (MIT Press)
[11] B M Adams, H T Banks, M Davidian, E S Rosenberg 2007 Estimation and Prediction with HIV-Treatment Interruption Data. *Bulletin of Mathematical Biology* **69** 563-584
[12] B M Adams, H T Banks, M Davidian, H D Kwon, H T Tran, S N Wynne, E S Rosenberg 2005 HIV Dynamics: Modeling data analysis and optimal treatment protocols *J. Comput. Appl. Math.* **184** 10-49
[13] Y Yang and X Xiao 2010 Threshold dynamics for an HIV model in periodic environments. *J. Math Analysis and Applications*, **361** 59-68
[14] A Jarne, D Commenges, M Prague, Y Levy, R Thiebaut Modeling CD4+T cells dynamics in HIV infected patients receiving repeated cycles of exogenous Interleukin 7 arXiv:1602.05399.